
Contents

<i>List of Figures</i>	<i>page</i> x
<i>List of Tables</i>	xiii
<i>List of Examples</i>	xv
<i>Preface</i>	xix
1 Introduction to GPU Kernels and Hardware	1
1.1 Background	1
1.2 First CUDA Example	2
1.3 CPU Architecture	10
1.4 CPU Compute Power	11
1.5 CPU Memory Management: Latency Hiding Using Caches	12
1.6 CPU: Parallel Instruction Set	13
1.7 GPU Architecture	14
1.8 Pascal Architecture	15
1.9 GPU Memory Types	16
1.10 Warps and Waves	18
1.11 Blocks and Grids	19
1.12 Occupancy	20
2 Thinking and Coding in Parallel	22
2.1 Flynn's Taxonomy	22
2.2 Kernel Call Syntax	30
2.3 3D Kernel Launches	31
2.4 Latency Hiding and Occupancy	37
2.5 Parallel Patterns	39
2.6 Parallel Reduce	40
2.7 Shared Memory	51
2.8 Matrix Multiplication	53
2.9 Tiled Matrix Multiplication	61
2.10 BLAS	65
3 Warps and Cooperative Groups	72
3.1 CUDA Objects in Cooperative Groups	75
3.2 Tiled Partitions	80

3.3	Vector Loading	85
3.4	Warp-Level Intrinsic Functions and Sub-warps	89
3.5	Thread Divergence and Synchronisation	90
3.6	Avoiding Deadlock	92
3.7	Coalesced Groups	96
3.8	HPC Features	103
4	Parallel Stencils	106
4.1	2D Stencils	106
4.2	Cascaded Calculation of 2D Stencils	118
4.3	3D Stencils	123
4.4	Digital Image Processing	126
4.5	Sobel Filter	134
4.6	Median Filter	135
5	Textures	142
5.1	Image Interpolation	143
5.2	GPU Textures	144
5.3	Image Rotation	146
5.4	The Lerp Function	147
5.5	Texture Hardware	151
5.6	Colour Images	156
5.7	Viewing Images	157
5.8	Affine Transformations of Volumetric Images	161
5.9	3D Image Registration	167
5.10	Image Registration Results	175
6	Monte Carlo Applications	178
6.1	Introduction	178
6.2	The cuRAND Library	185
6.3	Generating Other Distributions	196
6.4	Ising Model	198
7	Concurrency Using CUDA Streams and Events	209
7.1	Concurrent Kernel Execution	209
7.2	CUDA Pipeline Example	211
7.3	Thrust and cudaDeviceReset	215
7.4	Results from the Pipeline Example	216
7.5	CUDA Events	218
7.6	Disk Overheads	225
7.7	CUDA Graphs	233
8	Application to PET Scanners	239
8.1	Introduction to PET	239
8.2	Data Storage and Definition of Scanner Geometry	241
8.3	Simulating a PET Scanner	247

Contents

ix

8.4	Building the System Matrix	259
8.5	PET Reconstruction	262
8.6	Results	266
8.7	Implementation of OSEM	268
8.8	Depth of Interaction (DOI)	270
8.9	PET Results Using DOI	273
8.10	Block Detectors	274
8.11	Richardson–Lucy Image Deblurring	286
9	Scaling Up	293
9.1	GPU Selection	295
9.2	CUDA Unified Virtual Addressing (UVA)	298
9.3	Peer-to-Peer Access in CUDA	299
9.4	CUDA Zero-Copy Memory	301
9.5	Unified Memory (UM)	302
9.6	A Brief Introduction to MPI	313
10	Tools for Profiling and Debugging	325
10.1	The gpulog Example	325
10.2	Profiling with nvprof	330
10.3	Profiling with the NVIDIA Visual Profiler (NVVP)	333
10.4	Nsight Systems	336
10.5	Nsight Compute	338
10.6	Nsight Compute Sections	339
10.7	Debugging with Printf	347
10.8	Debugging with Microsoft Visual Studio	349
10.9	Debugging Kernel Code	352
10.10	Memory Checking	354
11	Tensor Cores	358
11.1	Tensor Cores and FP16	358
11.2	Warp Matrix Functions	360
11.3	Supported Data Types	365
11.4	Tensor Core Reduction	366
11.5	Conclusion	371
	<i>Appendix A A Brief History of CUDA</i>	373
	<i>Appendix B Atomic Operations</i>	382
	<i>Appendix C The NVCC Compiler</i>	387
	<i>Appendix D AVX and the Intel Compiler</i>	393
	<i>Appendix E Number Formats</i>	402
	<i>Appendix F CUDA Documentation and Libraries</i>	406
	<i>Appendix G The CX Header Files</i>	410
	<i>Appendix H AI and Python</i>	435
	<i>Appendix I Topics in C++</i>	438
	<i>Index</i>	448